Short Synopsis

For

Ph. D. Programme 2010-11

Title: Design and Development of Efficient Classification Techniques in Data Mining

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

Submitted by:

Name: Veena Mittal

Registration No.: 10019990291

Supervisor :

Name: Dr. Indu Kashyap

Designation: Associate Professor

Abstract

The current and previous research is done in static environments in case of data streams but there is a need of real time data streaming due to the potentially larger number of applications are available in the scientific and business analysis. The online learning is done in order to work in dynamic environments in which concept tend to change with time and the accuracy of classifiers decreases. There are several methods used in learning in presence of dynamic environments such as single classifier method in which only one classifier is kept at a time which is replaced with new one upon detection of changes in concepts. It is done because a significant increase in error rate is observed upon the change in concept. The traditional learners in single classifier approaches include neural networks, nearest-neighbor, decision rules, Naïve Baysian, etc. They are not able to adapt to changes but can process the data sequentially. The ensembles of classifiers are used to boost the accuracy of classifiers that includes bagging and boosting techniques. The boost in accuracy is achieved due to the modularity of ensembles which results in adapting to changes in a more natural way and less modification is needed. The diversity in ensemble members gives a more boost in accuracy.

The author will propose algorithm that is based on ensembles that ensures more accuracy in the classification in the data streams and also able to adapt to the changes in the target variable without modification. It will be done using the real world datasets and artificial datasets that are considered as benchmarks. The proposed algorithm will also be compared with the existing algorithms in order to assure the accuracy of the proposed method.

Keywords:- Data Streams, dynamic environments, classifiers, ensemble learning, online methods

CONTENTS

S.No	Description	Page No.
1	Introduction	1-8
	1.1 Motivation	1
	1.2 Data Mining	2
	1.3 Data Mining Techniques	3-8
	1.3.1 Association Rules	3
	1.3.2 Classification	4-6
	1.3.3 Clustering	6-7
	1.3.4 Applications of data mining	7-8
2.	Literature Review	9-11
	2.1 Drift Detection Methods	9
	2.2 Windowing Techniques	10
	2.3 Ensembles	11
3.	Description of Broad Area	12-19
	3.1 Data Streams	12
	3.2 Data Stream mining	13
	3.3 Introduction to data stream mining algorithm	14-19
	3.3.1 Drift Detector Methods	14
	3.3.2 Windowing Techniques	15
	3.3.3 Ensemble Algorithm	16-18
	3.4 Challenges of Big data	19
4.	Objectives of the study	20
5	Methodology to be adopted	21
6.	Expected outcome of the research	21
7.	References	22-24

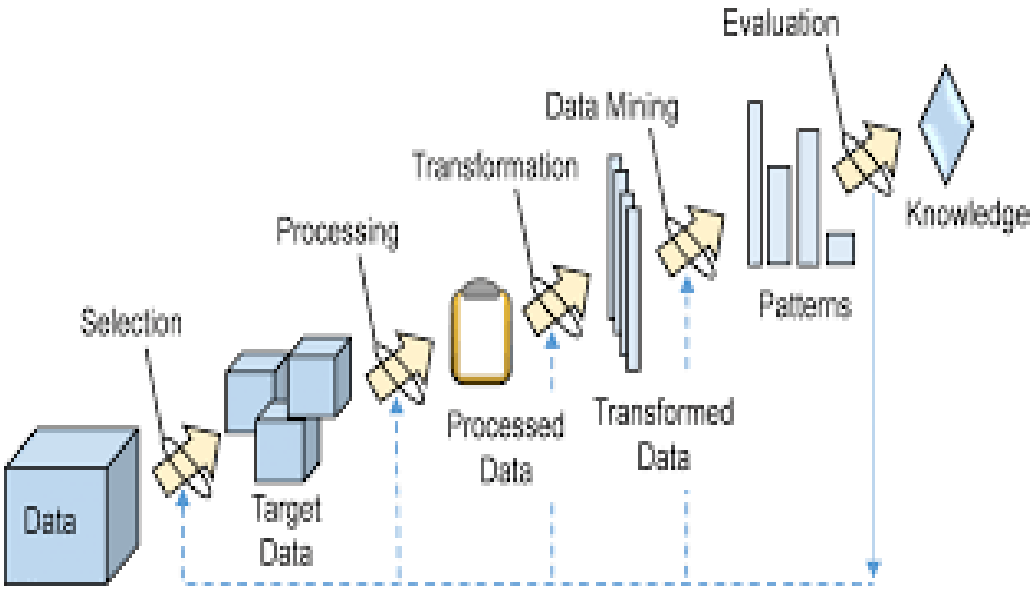
1.Introduction

1.1 Motivation

In the era of information technology, the data can be gathered and shared at anytime and anywhere in the world. It is implemented in many application areas like social networking, credit applications, fraud detection, manufacturing, medicine, banking, health care, research, stock market, entertainment, telecommunications, retail, finance banks and credit applications, etc. The machines learning which is a field of computer science gives the ability to machines to learn without being programmed for it and therefore it can be used to make predictions based on the learning of algorithms. A huge amount of data gathered from the storage in computer systems is presented in the form of datasets on which these learning algorithms are implemented. The datasets may contain many useful and interesting hidden patterns which can be extracted as knowledge, therefore the knowledge discovery methods are of great importance in data mining. These methods provide understandable patterns from very large data sets. The data mining techniques includes association rules, clustering and classification techniques. An enormous amount of data is available to decision makers that is not only large in size but also having a large variety and velocity which is difficult to handle using traditional methods of data mining. But there is a need to handle such data in order to extract knowledge out of them because such type of data is continuously increasing due to social area networking, we can say that today is an era of big data. The processing of such data is very challenging as the data is sourced from multiple sources and also it is having very complex and evolutionary relationship. The example of Bigdata is presidential debate between president Barack Obama & Governor Mitt Romedy triggered more than 10 billion tweets in just two hours. It is how fast the rate on which data is expanding. There are many more example available in these tweets, the moments that generated discussions of public interests related to Medicare and vouchers. It provides new means to judge the public opinion & generates feedback in real time which is more valuable. The speed and size of datasets make them impossible to store permanently in memory for which we have to employ scheme of forgetting mechanism. It is very challenging to deal with such data and it opens the door in research for mining such class of data as current research is dedicated to only static environments.

1.2 Data Mining

Data mining is the process of discovering patterns in very large data sets that involves methods at the intersection of machine learning, database systems, etc. It is the application of techniques (machine-learning) to large databases and it is a subfield of computer science. The goal of the data mining process is to extract useful information from the raw data(in the form of datasets) and transform it into an understandable structure for further use. It involves data management aspects, data pre-processing, complexity considerations. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. It has various stages that includes selection, pre-processing, transformation, mining and interpretations, etc.[1].



Steps in data mining in the discovery of knowledge

There are a sequence of steps that includes:

Data cleaning: Removing of noise and inconsistent data

Data integration: To combine data from multiple sources

Data Selection: The data which is considered as relevant for a particular task is retrieved from database

Data transformation: The applications of operations such as aggregations, summary to transform it into consolidated forms.

Data mining: The data patterns are extracted by the applications of intelligent methods.

Knowledge presentation: The mined knowledge is presented to the user in the suitable presentation techniques.

There are various issues in the mining of data such as *requirement of different types of knowledge to different users*, *interactive mining of knowledge* as it is difficult to know what exactly can be extracted from database, *incorporation of background knowledge* at various levels of abstraction, presentation of results in high level language so that can be understood by humans easily, it can be in the form of tables, graphs, charts, trees, etc. The *handling of noisy data* needs to be addressed as the regularities in data can cause confusion. The *evaluation of patterns* give thousands of patterns to uncover that represents the common knowledge.

The issues related with performance such as *efficiency and scalability of data mining algorithms*, development of *parallel and distributed mining algorithms* due to the computational complexity of data mining methods.

The handling of various complex data objects that includes spatial data, temporal data, hypertext ,multimedia databases, transaction data requires different mining of data of various kinds. The mining of data from *heterogeneous databases* requires connection from various sources of data that requires huge, distributed and heterogeneous databases. It discovers the knowledge in various forms which proves to be great challenge in data mining.

1.3 Data Mining Techniques

1.3.1 Association Rule Analysis[1]

It is used to find the associations and correlation in a large data set which is formed by collecting from the data which is continuously generated in many industries. The interesting patterns found from such data has proven very effective in decisions making processes, designing of catalogues and cross marketing. A

market basket analysis is a popular example of association rule mining,in which the buying habits of customer are analyzed and the associations are found between the items brought by customers in order to develop the market strategies.
Rule **support** and **confidence** are the measures for interestingness. It can be described as follows:
Support(A implies B) = P(A union B)
It means support is the percentage that contains both A and B.

Confidence(A implies B)= P(B|A)
It is the conditional probability, the percentage of containing A that also contains B.

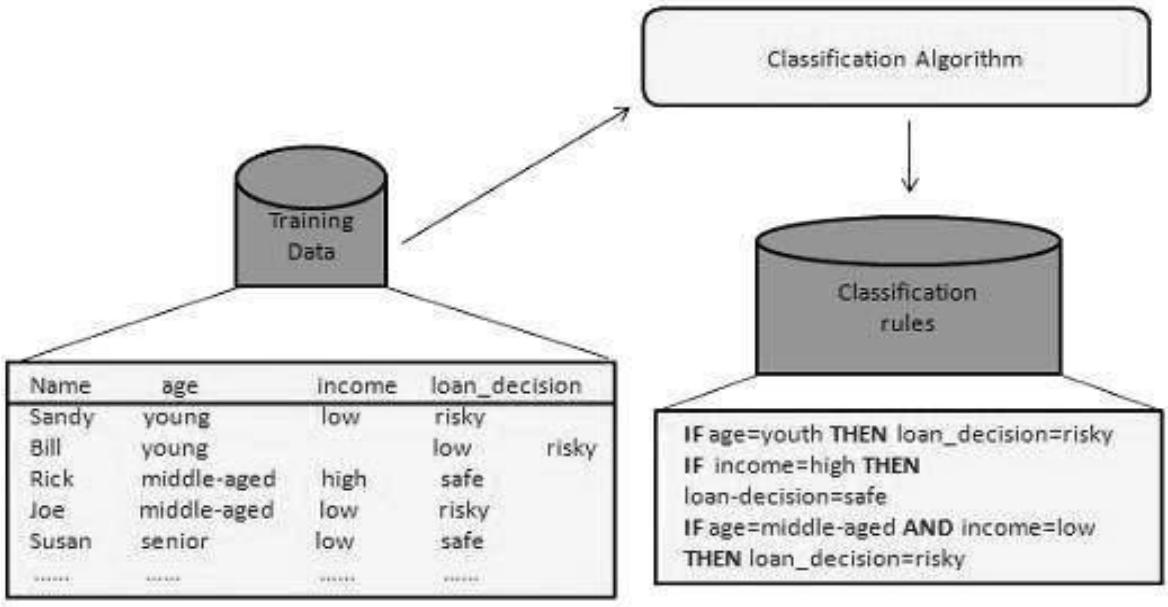
For example:
P(chips| beer) = 0.8

Then, rule can be defined as:
80 percent of customers who buy beer also buy chips.

P(BIA, D) it is the conditional probability that when A is happening , B also happens and D are the customer attributes which can be gender, marital status, age , income status, etc. This association can be used for targeting such customers and plan the market strategy accordingly. For example: some discounts can be offered to such customers in order to increase the sale.

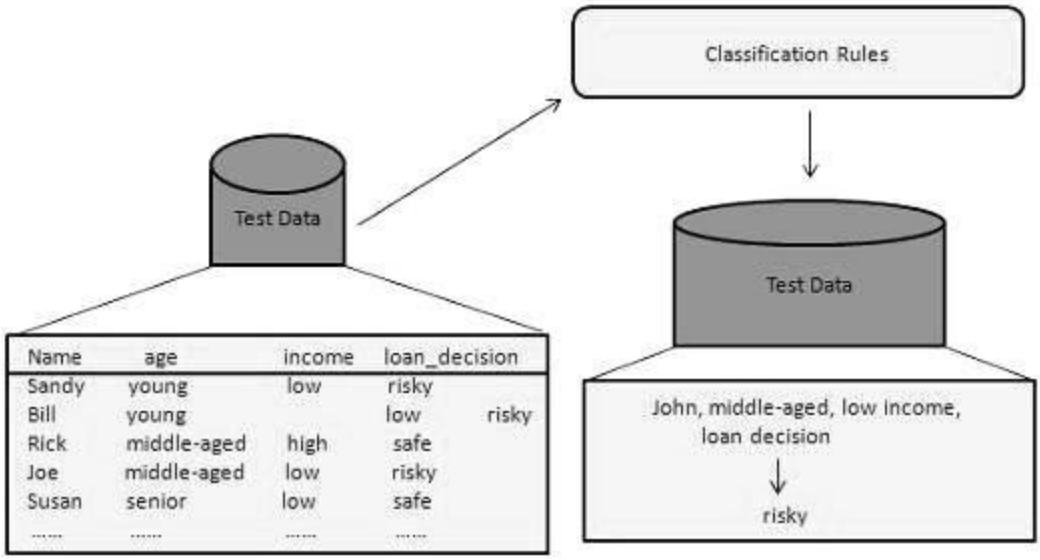
1.3.2 Classification

`This technique is used to predict future data trends and also for describing the data classes. It is also used for decision making. The model of classification can be used for categorizing the loan applications of banks whether they are safe or risky. It is a two step process in which a model is built in the first step that describes a set of data classes. It is constructed from the tuples of database. The tuples are then analyzed for the building of model which is called as data set. There are various criteria on which classification models are evaluated such as predictive accuracy, speed, robustness, scalability, interpretability, etc. There are various techniques used in classification such as Bayesian classification, decision tree induction, neural networks, etc.[1]



Building of Classification model(first step)

The rules of classification is stated using the IF and THEN condition. For example: training data consists of attributes such as name ,age, income and loan_decision. The classification rules are then applied on test data whose classification is to be made whether the loan decision is risky or safe.



Classification of data (second step)

1.3.3 Clustering

It is a group in which similar data is grouped into the same cluster. The data within the same cluster has high similarity as compared to data in the different clusters. It is different from the classification in the sense that we don't know about the class labels of each object. The dissimilarity is computed based upon different measures like distance, attributes, variable types, etc. The analysis of cluster is used in various applications such as data analysis, pattern recognition, market research, image processing, etc.[1]

There are various requirements for forming the clusters in data mining

1. **Scalability:** The algorithms work well when size of data sets is small but in case of large databases, results can be biased and highly scalable clustering algorithms of clustering are needed.
2. **Ability to deal with noisy data:** The presence of outliers in real world databases may produce incorrect clusters.
3. **Constrained based clustering:** To perform clustering in presence of various constraints in real world applications.
4. **Ability to deal with different types of attributes:** There are various types of data such as categorical, ordinal data and mixtures of these data.

There are several types of clustering techniques

1. Partitioning methods
2. Density based methods
3. Hierarchical methods
4. Grid-based methods

Partitioning methods

A database of n objects is given and clusters have to be formed from them. In this method, k partitions are constructed from the data such that $k<n$ or $k=n$ but it must satisfy two conditions , each object should contain only in one cluster and each object should have cluster, no data element should be left without being in the cluster. The iterations are made in order to move from one cluster to another based upon the distance. It is called as **iterative relocation technique.** The classical partitioning methods include k-means and k-medoids[1].

Density based methods

The partitioning method is based upon distance which is to be measured between the objects. But it faces difficulty in arbitrary shapes of clusters other than spherical shape. This method is based upon the density in which the cluster will grow as long as density exceeds some threshold in the neighborhood. It is particularly used in the discovering clusters of arbitrary shape and in filtering of noise. **DBSCAN** is the clustering method which is density based that grows in size according to threshold of density.

Hierarichical methods

It creates a hierarchical decomposition from the given set of data objects. It can be classified as divisive or agglomerative. The divisive approach is top-down which starts with assumption that all the objects are in same cluster and it is spilt up into smaller clusters in each of the successive iterations until the termination condition holds. The agglomerative approach is bottom-up in which each object is considered in separate group and merges successively until termination.

Grid-based methods

The object space is quantized into a finite number of cells that forms a grid structure and all the operations related to clustering are performed on this grid. The boundary of all the clusters formed is either vertical or horizontal not the diagonal. The major advantage of this approach is the fast processing time but clusters are formed of low quality due to isothetic nature of algorithm. There are two approaches that are both density based and grid-based named CLIQUE and Wave Cluster. The typical example of grid-based method is STING.

1.4 **Applications in data mining**

The **applications** of data mining in various fields such as DNA data analysis, financial data analysis, retail industry, telecommunication industry and various business sectors such as risk management and targeted marketing[1].

In **DNA data analysis**, various interesting sequential pattern analysis and also the search techniques related with similarity can be done with the help of data mining techniques. It includes semantic distribution of heterogeneous and distributed

database, association analysis which identifies the co-occurring of gene sequences and genetic data analysis including the complex structures presented in the form of graphs, trees, cuboids, etc.

In **financial data analysis,** systematic data analysis can be done using data mining techniques such as viewing of changes in debt and revenue by month, sector, region, etc. To predict the payment of loan and analysis of customer credit policy, it is very important application in the business of banks. The detection of various financial crimes such as money laundering can also be done with the help of data mining. It can also be used to classify the customers for targeted marketing.

In **retail industry,** it is considered as one of the major application area in data mining due to the expansion of e-commerce industry in which large amount of data is available in the form of shopping history, service records, transportation of goods, etc. which helps in predicting the buying behavior of customer, shopping trends and patterns, retention of better customer. It can also be used to analyze the effectiveness of sales campaigns which helps in improving the profits of company. The multidimensional analysis of sales, trends and fashions including cost, profit and service of commodities is also done with the help of mining techniques. The mining associations can be used to suggest the cross-reference of items for the recommendations of customers.

In **telecommunication industry,** the multidimensional analysis such as calling-time, type of call ,etc along with system load, resource usage, profit can be performed. The analysis of fraudulent pattern saves the millions of dollars in a year. The promotional services of the sector can also be designed by finding associations in the calling pattern of customers. The usage of visualization tools proves to be very effective in this industry.

In **business applications** such as insurance relies on analysis of data to make the maximum profit. The overcharging of low-risk and undercharging of high-risk policyholders would decrease profits, therefore effective analysis is required for the creation of accurate model. Similarly in targeted marketing, identification of potential customers can be very profitable to the business and data mining activities help in such identification.

.

2.Literature Review

The Literature Review is divided into three sections. Section 2.1 deals with the attempts that are made in the literature to deal with data streams using single classifiers methods called as drift detectors methods. Section 2.2 discuses the work done for the mining of data streams using windowing techniques. Section 2.3 discusses the classification in data stream mining using the ensemble approach.

2.1 Contribution of authors for data stream mining using the drift detector methods is discussed below:

Gama et.al[5] proposed a method called drift detection method . It is assumed that error rate of online classifier decreases with advancement of time if the target concept is stationary. But error rate will increase significantly in case of changing concept. There are two thresholds in DDM , one for warning level and another one is for drift level. In warning level it stores the examples in a short-term memory for the rebuilding of online classifier. On the other hand upon reaching of drift level, DDM reinitializes all the variables. The advantage of this approach is that it works well for in the detection of sudden and significant changes. The disadvantage of this approach is that it is not capable in detecting small and gradual changes.

Baena-Garc et.al[4] proposed a method called early drift detection method(EDDM) . It is an improved form of DDM to detect small and gradual changes. It considers the distance between the occurrence of two classification errors. When this distance decreases significantly, then it assumes the occurrence of drift. It uses the measure of average distance calculation between two errors and this measure is more sensitive towards gradual changes as compared to DDM. Like DDM, it also stores examples in short term memory upon reaching of warning level and rebuilds an online classifier upon reaching the pre-specified drift level.

Nishida and Yamauchi[3] proposed an approach called STEPD. This method monitors the two predictive accuracies called recent accuracy and overall accuracy for the online classifier. It is based on the assumption that there is statistical significance between these two accuracies after the change in the concept of target class. Furthermore, the statistical test of equal proportions is used to compare these two accuracies. Like DDM and EDDM it is also having two levels : warning level and drift level and therefore examples are stored in short term memory till the

occurrence of warning condition and reinitializes all the variables of an online classifier upon reaching the drift level in the same manner as that of DDM and EDDM. This method works better in detection of gradual changes as compared to EDDM and much faster than DDM in case of sudden changes.

2.2 Contribution of authors for data stream mining using the windowing techniques is discussed below:

Cohen and Strauss[6] suggested the use of decay functions for the implementation of forgetting process and making it more dynamic. This technique is called as weighted windows in which lesser weights are assigned to older process and therefore will be treated as of less importance by base classifier. This method analyzed the use of decay functions of different types to calculation of data stream aggregates.

Bifet[7][8] proposed the concept of sliding window which is adaptive and the algorithm is called as ADWIN. It is suitable for the sudden changes in the data streams. The idea behind this algorithm is that it keeps a sliding window having the most recently read examples and when two sub-windows are having distinct enough averages then there can be a change between corresponding values and detection of sudden changes is considered, also, the portion of window which is older is dropped.

Zliobait[11] proposed a family of algorithms and it is called as FISH. The family consists of FISH1, FISH2 and FISH3. The family use the similarities of time and space between examples to create a window dynamically.

Klinkenberg et al. [9] and Tsymbal et al.[10] proposed two windowing methods that are incorporated by FISH2 and FISH3 in order to implement a window of variable sample size.

2.3 Contribution of authors for data stream mining using the ensemble algorithms is discussed below:

Street and Kim[12] proposed an ensemble algorithm called Streaming Ensemble Algorithm . It changes the structure of an algorithm so that it can react to changes. The idea of this algorithm is the replacement of weakest expert based on accuracy

and diversity. The accuracy is used because it is the most prime factor, that the most recent examples should be correctly classified in order to detect and adapt to changes. The diversity uses in bagging and boosting which are the techniques used to give the boost to ensembles.

Wang et al[14] proposed an algorithm called Accuracy Weighted Ensemble which is used for restructuring of ensembles. A new classifier is trained on the most recent data chunks which is then used to evaluate the existing ensemble members in order to select the best component classifiers.

Kirby[15] proposed an ensemble method which uses Hoeffding trees and the method is called as Hoeffding Option Tree. It is similar to the method named Option decision trees which is proposed by Kohavi and Kunz[ ]. In option decision tree, the decision path splits the tree into several sub-trees and also the training example updates a set of option nodes instead of a single leaf. It combines the prediction of all applicable leaves into one result while making a decision. The Hoeffding Option Trees(HOT) works as a set of weighted classifiers in the compact structure provided by it. Like Hoeffding Trees, HOT is also built in an incremental fashion.

Bifet et al.[17][18] also proposed an ensemble method called Adaptive-Size Hoeffding Tree Bagging(ASHT) and it is strictly designed for Hoeffding trees. It provides the diversity in the components of ensembles by using different sizes trees because the smaller trees are more adaptable to changes. It also provides the forgetting mechanism which is very useful in data stream mining.

Oza[13] proposed an online version of bagging and boosting techniques in which component classifiers combine their decision using a simple majority vote and these classifiers are in incremental fashion, therefore, they are also called as incremental learners.

D.Brezezinski and J.Stefanowski[21] proposed an algorithm called Accuracy Updated Ensemble(AUE). This algorithm is designed to react equally to all types of changes .The existing classifiers work well for any single type of change. Some of them are good for detection of gradual changes while other are good for sudden changes, most of them specialize for one type of change. In AUE, the accuracy based weighted mechanism from block-based ensembles are combined with the incremental nature of Hoeffding Trees in order to incorporate all types of changes.

Description of Broad Area

3.1 Data Stream

It is an sequence of instances, having very fast speed of arrival which is impossible to store in a memory permanently. It is very large in size, due to which, most of the data mining approaches fail to process.[19]

The constraints imposed by data streams are as follows:

1. The information are stored in memory in the form of small summaries to carry out computations and rest of the information is discarded as it is impossible to store each and every part of data from data streams. It imposes the constraint for limited usage of memory.

2. The elements from data streams should be processed in real time due to their fast arrival speed. Therefore, online learning is a need for data stream mining. It imposes the time constraint in which processing is to be done.

3. The distribution which is generated over data streams tend to change with time, which may be irrelevant and proves harmful for current memory. It imposes the constraint to study the learning of evolving data streams rather than static learning algorithms.

3.2 Data Stream mining

Online learning

It is also termed as incremental learning and it focuses to process the incoming data from data streams sequentially and also takes care about the accuracy of the trained classifier in such a manner, as if, it has been trained on the whole data set. The classifier is updated in order to accommodate new training point[19].

An online classifier should possess the following qualities:

1. **Incremental:** The algorithm should be able to read blocks of data at a time, as the whole data is not available at the beginning.

2. **Single pass:** The data is available for only one time to the algorithm

3. **Limited time and memory:** The arrival speed of data is very high and it cannot be stored permanently in the memory so it should be processed within limited time and memory.

4. **Any-time learning:** The algorithm should give the best possible answer even when stopped in between without its conclusion

Forgetting mechanisms

A classifier in data stream should be able to respond to the evolving data. It can be achieved by forgetting of outdated data and it is difficult to select the data range which is to be remembered. The solution of this problem is to forget the training objects at a constant rate and it can be achieved by using a window having latest examples and used to train the classifier. The size of window is having tradeoff between stability and flexibility. The small size window will adapt to the changes quickly
but accuracy can be low due to insufficiency of data in the window. The large window size provides stability but may not be able to adapt to the changes quickly.

A decay function is used to differentiate the impact of data points.
This approach tries to give better results in order to maintain the balance between its accuracy and flexibility.

3.3 Introduction to data stream mining algorithms

3.3.1 Drift detection methods

This method is based on the fact that in each iteration, the decision class is predicted by using the online classifier. The number of classification errors is predicted using the random variable from Bernoulli distribution. The standard deviation is calculated using the probability of false prediction.

Algorithm: Drift Detection methods
Input: S: a data stream of examples
C: classifier
Output: W: a window with examples selected to train classifier C
1: Initialize(i; pi; si; psmin; pmin; smin);
2: newDrift false;
3: W ;;
4: W0 ;;
5: **for all** examples xi 2 S **do**
6: **if** prediction C(xi) is incorrect **then**
7: pi pi + (1:0 □ pi)=i;
8: **else**
9: pi pi □ (pi)=i;
10: compute si using the formula of standard deviation
11: i i + 1;
12: **if** i > 30 (approximated normal distribution) **then**
13: **if** pi + si _ psmin **then**
14: pmin pi;
15: smin si;
16: psmin pi + si;
17: **if** drift detected using the formula for alarm level **then**
18: Initialize(i; pi; si; psmin; pmin; smin);
19: W W0;
20: W0 ;;
21: **else if** warning level reached **then**

22: **if** newDrif t = true **then**
23: W0 ;;
24: newDrif t false
25: W0 W0 [fxig
26: **else**
27: newDrif t true;

3.3.2 Windowing Techniques

The sliding windows are used in dealing with time changing data by limiting the size of size of window from which the old concept is eliminated . It should have the property of any time learning and able to provide best answer at any moment in the processing of the input stream. The forgetting process is the key element of windowing techniques. The oldest concept that is not fitting in the window is discarded but includes the most recent examples.

Algorithm The basic windowing algorithm

Input: S: a data stream of examples
W: window of examples
Output: C: a classifier built on the data in window W
1: initialize window W;
2: **for all** examples xi 2 S **do**
3: W W [fxig;
4: if necessary remove outdated examples from W;
5: rebuild/update C using W;

3.3.3 Ensemble methods

The algorithm of ensemble is a set of components in which decision is formed by the aggregation of voting rule. The decision of ensemble is more accurate as compared to single classifier.

The diversity in the components of ensembles can be made based upon the three parameters:

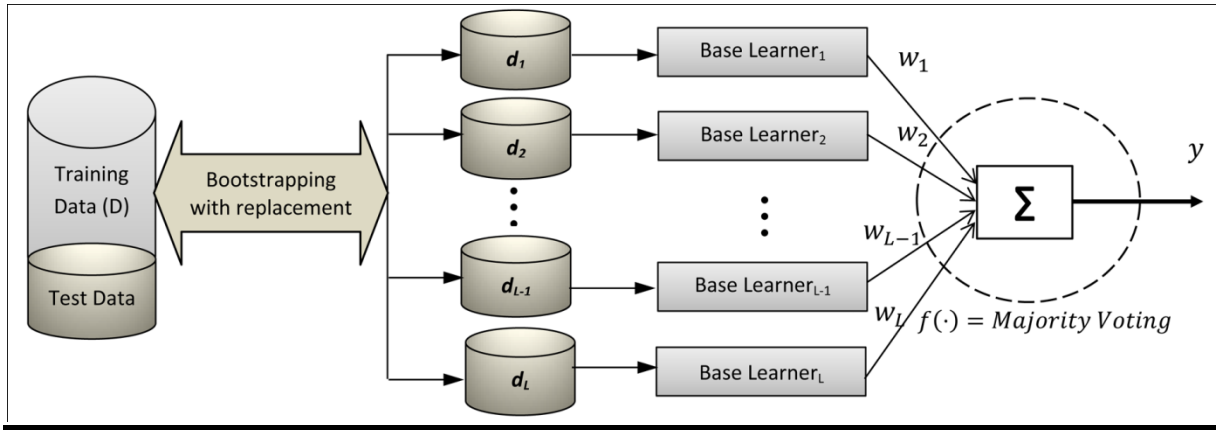
1. Attributes used by the component
2. Training data used for training of components
3. Base learner from which they are created

Algorithm Generic ensemble training algorithm [51]
Input: S: a set of examples
k: number of classifiers in ensemble
Output: E: an ensemble of classifiers
1: E k classifiers;
2: **for all** classifiers Ci in ensemble E **do**
3: assign a weight to each example in S to create weight distribution Dm;
4: build/update Ci with S modified by weight distribution Dm;

There are two techniques which are used to increase the accuracy of the classifiers of ensembles named **Bagging** and **Boosting**

Bagging

It is a voting method in which base learners are trained over different training data. It generates slightly different samples from a given sample of size N. The samples can be similar as they are taken from the same sample but can be different due to random replacement of samples. The learning algorithm is called unstable algorithm when small changes made in the training set and it gives more impact in the generated learner. In bootstrap aggregating, it generates L(L<N) base learners using this unstable algorithm, averaging is done during testing. This method is applicable to regression also but instead of average, median is used at the combining the predictions.



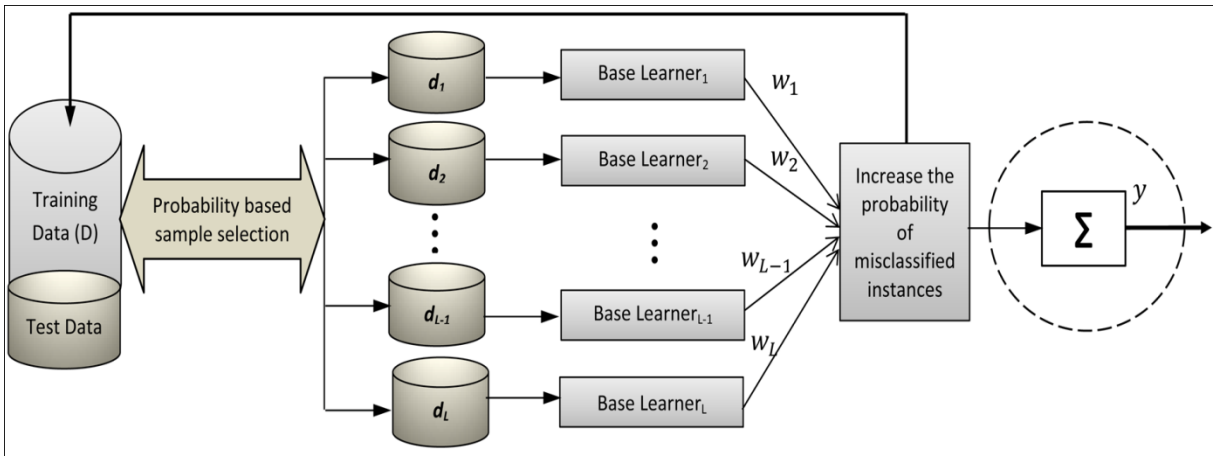
Bootstrap with Replacement

Boosting

In this technique, the next learner is trained on the mistakes (misclassification) of the previous learners in order to generate complimentary base learners. It combines three weak learners and generates one strong learner which is having a small error probability. The disadvantage of this approach is that it requires a large training set which is divided into three partitions randomly. The first classifier is trained on first partition of dataset. It takes all the instances which first classifier misclassifies for the second partition and many instances which are correctly classified will together form the dataset for second partition on the next classifier is trained. The third partition data is exposed to both the previous two classifiers and all the misclassified data using both of them will become the data set for third classifier. In the testing phase, an instance is given to first two classifiers and if misclassified then the classification from the third one is considered as output. In case of small size of datasets, the second and third classifiers will be trained only a subset of dataset which are misclassified by first classifier.

In adaptive boosting, there is an algorithm named AdaBoost, which can combine any number of base-learners not necessary to be three and it overcomes the problem of large dataset because the same training data is used again and again. The technique used by AdaBoost is to make modifications in the probabilities of drawing the instances which is considered as function of error.

It can be presented diagrammatically as follows:



Probability based sample selection

The algorithm for AdaBoost is given below:

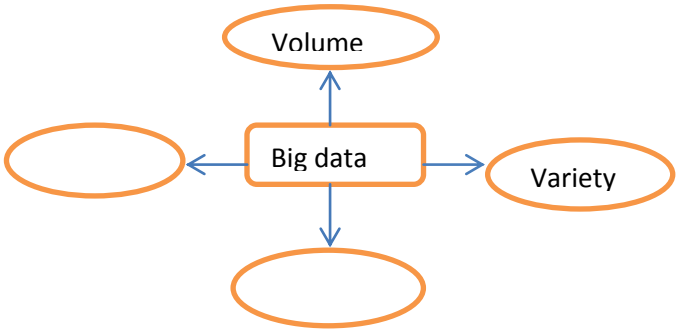
Training:
For all {xt, rt}~l E *X,* initialize p~ = *liN*
For all base-learners j = 1, ... ,*L*
Randomly draw *Xj* from *X* with probabilities *P)*
Train *dj* using *Xj*
For each *(xt,rt),* calculate *y5 - dj(xt)*
Calculate error rate: fj - It *P) . l(y5 f= r t)*
If fj > 1/2, then *L* - j - 1; stop
I3j - fj/O - fj)
For each *(x t , r t),* decrease probabilities if correct:
If *Yjt = r t Ptj+l - 13jPtj* EI se *Pij+l - Ptj*
Normalize probabilities:
Zj - It *P)+l;* P)+l - P)+l / *Zj*
Testing:
Given *x,* calculate *dj(x),j = 1, ,L*
Calculate class outputs, i = 1, , *K:*
Yi = I]=l (log *JJ) dji(x)*

Figure 15.2 AdaBoost algorithm

3.4 Challenges with Big data mining

The challenges of big data are manifold. It focuses on how to access data and how to apply computational operations on it related with arithmetic computing. The big data storage location is not single but sparse at different places and the characteristics of big data is that it is continuously growing in nature, therefore, there is a requirement of distributed data storage at large scale. It is also expensive to move data across different locations. It is required in case of data mining algorithms, the data, to be loaded in main memory for computing purposes. The challenges are also related with semantics & domain knowledge of big data[2]

The challenges also focuses on designing of algorithms in handling the difficulties of complex and dynamic data. The sharing of information is required for better development and processing at each stage. The challenges of big data can also be studied by considering the four V of big data i.e. volume, velocity, variety, veracity as shown in Fig 1 The volume is increasing continuously in the form of click stream, log, event, speech, social media etc. The velocity is speed of generation of data, it can also be called as rate of analysis. The variety deals with different types of data like structured, unstructured and semi-structured. The veracity is data in doubt, distrusted and unclean. The uncertainty is due to data inconsistency and incompleteness of data[2].



Challenges of big data

4. Objectives

- To compute models over data streams, which can accommodate changing environments from where the data is drawn and it is very complex in the context of large streaming data.
- To design classifiers that are able to handle ultra-high dimensional classification problems along with high speed data streams.
- To handle mining of data that is continuous rather than an one-shot process in which existing techniques of data mining is still too slow to handle data.
- To study and implement the methods of data mining, beyond the process of classification so that high speed data streams can be handled.
- To work in the newest applications such as sensor networks, monitoring, telecommunication, or web log analysis in which learning is done in dynamic environments and data is also generated continuously. Furthermore, it also implies constraints such as memory usage, restricted learning, and testing time.
- To study and analyze the impact of change in the concepts with respect to time due to non-stationary nature of data streams which deteriorates the accuracy of classifier.
- To address the issues regarding the forgetting of outdated data and also the quick adaptation to most recent states in which simple incremental learning is not sufficient to deal with.

5. Methodology To Be Adopted

1.The algorithms will be proposed based upon ensemble methods which take care of learning in presence of drifts. It improves the false detection of drifts.

2.The various performance measures are used such as prequential accuracy, Kappa statistics, CPU time for the evaluation of the proposed algorithm. The proposed algorithm is compared with the state –of-the-art algorithms.

3.An ensemble will be designed that works well with various types of drifts such as sudden, gradual, abrupt, gradual, etc. It updates the classifier according to the current trend by the application of forgetting mechanism which classifies the data correctly.

4.The proposed algorithm will be implemented on various real time datasets available in the UCI repository and synthetic datasets which are used as benchmarks for evaluation purpose.

6.Expected Ootcome Of The Research

The new algorithm will be designed based on the accuracy weighted ensembles, Hoeffding Option Tree, Streaming ensemble algorithm , accuracy diversified ensembles that incorporates the forgetting mechanism which is the primary requirement of data stream mining and also the adaptation to the most recent states very quickly. The proposed algorithm removes the drawback of instability which is caused due to the smaller window size in windowing techniques and it predicts the concept which tend to change with time incorrectly. The proposed technique will come up to work with more accuracy so that classification of data of will be done more accurately.

7. References:

1.Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

2. X. Wu, X. Zhu, G. Wu and W. Ding , "Data Mining with big data" IEEE Trans. Knowledge and Data Eng.,. 26, no. 1, pp. 97-106, Jan. 2014

3. K. Nishida and K. Yamauchi, "Detecting Concept Drift Using Statistical Testing," in Proc. of the 10th International Conference on Discovery Science, vol. 755, 2007, pp. 264–269.

4.M. Baena-Garcıa, J. Del Campo-Avila, R. Fidalgo, and A. Bifet, "Early Drift Detection Method," in *Proc. of the 4th ECML PKDD International Workshop on Knowledge Discovery from Data Streams*, pp. 77-86, 2006.

5. J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with Drift Detection," in *Proc. of the 17th Brazilian Symposium on Artificial Intelligence (SBIA'04)*, pp. 286-295, 2004.

6. Edith Cohen and Martin J. Strauss. Maintaining time-decaying stream aggregates. *J. Algorithms*, 59(1):19–36, 2006.

7. Albert Bifet and Ricard Gavald`a. Kalman filters and adaptive windows for learning in data streams. In Ljupco Todorovski, Nada Lavrac, and Klaus P. Jantke, editors, *Discovery Science*,volume 4265 of *Lecture Notes in Computer Science*, pages 29–40. Springer, 2006

.

8. Albert Bifet and Ricard Gavald`a. Learning from time-changing data with adaptive windowing.In *SDM*. SIAM, 2007.

9. Ralf Klinkenberg and Thorsten Joachims. Detecting concept drift with support vector machines.In Pat Langley, editor, *ICML*, pages 487–494. Morgan Kaufmann, 2000.

10. Alexey Tsymbal, Mykola Pechenizkiy, Padraig Cunningham, and Seppo Puuronen. Dynamic integration of classifiers for handling concept drift. *Information Fusion*, 9(1):56–68, 2008.

11. Indr˙e ˇZliobait˙e. *Adaptive training set formation*. PhD thesis, Vilnius University, 2010.

12. W. Nick Street and YongSeog Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *KDD*, pages 377–382, 2001.

13. Nikunj C. Oza. *Online Ensemble Learning*. PhD thesis, The University of California, Berkeley, CA, Sep 2001.

14. Haixun Wang, Wei Fan, Philip S. Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *KDD*, pages 226–235. ACM, 2003.

15. Richard Kirkby. *Improving Hoeffding Trees*. PhD thesis, Department of Computer Science,University of Waikato, 2007.

16. Ron Kohavi and Clayton Kunz. Option decision trees with majority votes. In Douglas H.Fisher, editor, *ICML*, pages 161–169. Morgan Kaufmann, 1997.

17. Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, and Ricard Gavald`a. Improving adaptive bagging methods for evolving data streams. In Zhi-Hua Zhou and Takashi Washio, editors,*ACML*, volume 5828 of *Lecture Notes in Computer Science*, pages 23–37. Springer, 2009.

18. Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, Richard Kirkby, and Ricard Gavald`a. New ensemble methods for evolving data streams. In John F. Elder IV, Fran¸coise Fogelman-Souli´e, Peter A. Flach, and Mohammed Javeed Zaki, editors, *KDD*, pages 139–148. ACM, 2009.

19. Albert Bifet. Adaptive learning and mining for data streams and frequent patterns. PhD thesis, Universitat Polit´ecnica de Catalunya, 2009.

20.K. Nishida, K. Yamauchi, and T. Omori, "ACE: Adaptive Classifiers-Ensemble System for Concept-Drifting Environments," *in Proc. of the 6th International Workshop on Multiple Classifier Systems*, vol. 3541, 2005, pp. 176–185

21. Brzeziński, Dariusz, and Jerzy Stefanowski. "Accuracy updated ensemble for data streams with concept drift." *International Conference on Hybrid Artificial Intelligence Systems*. Springer Berlin Heidelberg, 2011.

22. Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2014.